International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Forecasting Agricultural Yield and Market Price using Time Series Models

1st Shashikant Lohar
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India
shashilohar21@gmail.com

2nd Sakshi Patil
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India
sakshipatil2061@gmail.com

3rd Prathmesh Pawar
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India
prathmeshpawaru@gmail.com

4th Akshata Rahinj
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India
akshatarahinjpict25@gmail.com

5th Kimaya Urane
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India krurane@pict.edu

***Abstract*—Agriculture is a foundation of global food security, but it's vulnerable to changeable environmental conditions and shifting request prices. This design focuses on developing an advanced system for soothsaying agrarian yields and request prices using machine learning algorithms. By integrating data from various sources, similar as literal rainfall patterns, soil quality criteria and crop product records, the system aims to give accurate prognostications. These perceptivity will help growers and policymakers in optimizing agrarian practices and making informed opinions, eventually enhancing productiv- ity and profitable stability. The design contributes to the growing field of agrarian technology, addressing challenges related to food security and sustainable husbandry practices in the face of environmental and request misgivings.**

***Index Terms*—Agrarian soothsaying, Machine learning, Crop Yield vaticination, Market Price Prediction, Data Integration, Prophetic Modeling, Sustainability.**

I. Introduction

Agriculture plays a vital part in the global frugality and is integral to icing food security. still, the agri- artistic sector is vulnerable to several changeable factors similar as environ- mental conditions, shifting request prices, and soil variability, which significantly affect crop product and profitability. For countries like India, where husbandry sustains a large portion of the population, the capability to read agrarian yields and request prices is pivotal to mollifying pitfalls and optimizing resource allocation.

Advancements in Artificial Intelligence(AI) and Machine literacy(ML) have converted numerous diligence, and agricul- ture is no exception. AI- powered models offer the eventuality to dissect vast quantities of agrarian data, including literal rainfall records, soil characteristics, and request trends, to give prophetic perceptivity that can support decision- making for growers, dealers, and policymakers. While countries like the United States, Canada, and Australia have made signifi- cant strides in developing AI- grounded agrarian soothsaying models, the different and unique agrarian geography of India requires acclimatized results that can regard for the country's specific climate, soil types, and crop patterns.

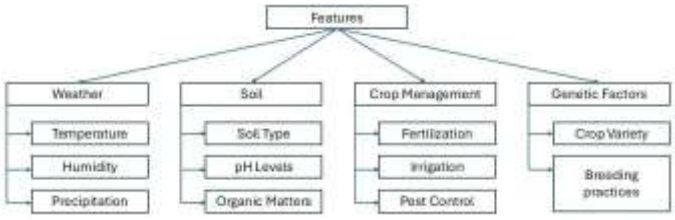

Fig. 1. Factors affecting on forecasting agricultural yields and prices.

A. Weather Conditions:

Weather is a critical factor in crop product, directly im- pacting factory growth and development. Temperature influ- ences the rate at which shops perform essential processes like photosynthesis and respiration, with each crop having an optimal temperature range for maximum growth. rush provides the necessary water for shops, enabling hydration and nutrient immersion from the soil, but too important or too little can lead to poor yields. moisture also plays a significant part by affecting how shops chance and their vulnerability to conditions. High moisture can foster the growth of dangerous fungi, while low moisture might beget shops to lose water and hamper their growth.

B. Soil Characteristics:

The quality and parcels of soil are abecedarian to how well crops can grow. Soil type dictates water retention and nutrient vacuity, with different soils offering colorful situations of support for factory growth. For illustration, flaxen soils

drain snappily, while complexion soils hold humidity. The pH position of soil influences which nutrients are available for factory uptake, and maintaining a balanced pH is important for nutrient immersion. Also, organic matter enriches the soil by perfecting its structure, water- holding capacity, and fertility, creating a better terrain for factory roots and salutary organisms, eventually leading to advanced crop productivity.

C. Crop Management Practices:

Successful crop operation is essential for optimizing yields and maintaining soil health. Proper fertilization replenishes nutrients that might be depleted in the soil, and choosing the right type and quantum of diseases can significantly enhance factory growth. Effective irrigation systems insure crops admit acceptable water, especially in areas with irregular rain- fall, reducing the threat of water stress. Also, effective pest and complaint control strategies are pivotal to minimize crop losses. ways like fungicide operation, crop gyration, and integrated pest operation help cover shops from damage, promoting healthier and further productive crops.

D. Genetic Factors:

The inheritable traits of crops largely determine their growth eventuality and adaptability. Different crop kinds come with unique inheritable features that affect their capability to yield, repel conditions, and acclimatize to environmental conditions. Advances in breeding practices have led to the creation of high- yield, complaint- resistant, and failure-tolerant crops. By fastening on the selection and development of these bettered kinds, growers can achieve better yields indeed under grueling conditions. This makes inheritable advancements a foundation of ultramodern husbandry, supporting sustainable crop product and enhancing food security.

The success of AI models in soothsaying agrarian yields and prices depends heavily on the quality and volume of available data. Integrating data from multiple sources, similar as satellite imagery, meteorological data, and agrarian checks, can help make more accurate and dependable prophetic models. still, the challenge lies in carrying and homogenizing this data, especially in regions where agrarian data collection is limited.

This design aims to address these challenges by developing a machine literacy- grounded system that forecasts agrarian yields and request prices. The system will dissect colorful data sources to induce prognostications, helping stakeholders optimize crop operation practices, plan crops, and better understand request trends.

II. LITERATURE SURVEY

In a study [1] concentrated on prognosticating crop yield in India, the authors employed colorful machine literacy models, in- cluding artificial neural networks(ANN), logistic retrogression, generalized direct models, direct discriminant analysis(LDA), support vector machine(SVM), and grade Boosting Tree. The datasets used include current diurnal commodity prices from colorful requests(Mandi),district-wise queries from the Kisan Call Centre(KCC), and crop-specific data similar as irrigated area and request advents. The study stressed the limitations of simplistic direct styles, championing for ensemble ap- proaches that synthesize multiple models to ameliorate crop yield prognostications, considering India's different agrarian conditions. In a study [2] on crop yield vaticination, Morales- Villalobos et al. explored the effect of colorful machine learning mod- monorails — formalized direct models, arbitrary timber, and artificial neural networks on prognosticating yields for sunflower and wheat in five regions of Spain. The dataset, named data- handwriting- morales- villalobos, was generated from biophysical crop mod- monorails(OilcropSun and Ceres- Wheat) using simulations of ranch- position data(2001 – 2020). The Random Forest model outperformed neural networks and direct models, showing a Root Mean Square Error(RMSE) of 35 – 38%. still, prognosti- cations had limited enhancement over birth pars, emphasizing the need for caution when applying machine literacy for yield soothsaying. In a study [3] on crop yield vaticination for Maharashtra, experimenters used machine literacy models similar as ANN, SVM, KNN, Decision Tree, Random timber, GBDT, and Formalized Greedy timber, with a focus on the Random Forest algorithm for indigenous yield vaticination. The dataset used was the Horticulture Area product Yield and Value for Spice Crop, incorporating five climatic parameters. The model was trained using 20 decision trees, achieving an delicacy of 87%. A10-foldcross-validation fashion bettered model trustability. This exploration underscores the effec- tiveness of machine literacy in soothsaying crop yields and informs analogous sweats in prognosticating yields and request prices. In another study [4] on crop product vaticination, the authors employed the Random Forest algorithm to estimate yields grounded on colorful attributes similar as state, quarter, crop time, season, crop type, area, and product. The dataset incorporated historical data to prognosticate crop yields more directly, addressing challenges like rainfall, water vacuity, and soil quality. The model aimed to help growers in making informed civilization opinions. By using machine literacy ways, this study emphasizes the significance of prophetic ana- lytics in husbandry, with bettered model evaluation parameters like delicacy and perfection driving the results. In a study

[5] concentrated on crop yield vaticination, machine literacy models similar as Back Propagation Neural Network(BPNN), Support Vector Machine(SVM), and General Regression Neu- ral Net- work(GRNN) were employed. The dataset, collected from colorful agrarian departments and meteorological centers in Tamil Nadu, included attributes like downfall, evapotran- spiration, rush, temperature, and toxin use over an 18 time period. The GRNN model outperformed others, achieving a 97% delicacy(R^2 = 0.97) with a normalized mean square error of 0.03, pressing its efficacity in prognosticating crop yields across different geographical fields.This paper [6] reviews the integration of machine literacy and statistical ways for crop yield vaticination, pressing the significance of data- driven approaches in husbandry. It evaluates models similar as Bayesian spatial generalized direct models, retrogression analysis, and machine literacy algorithms (e.g., Random For-

est, XGBoost) using different datasets from government and meteorological sources. crucial performance criteria , including delicacy, recall, perfection, and F- score, are bandied to assess prophetic capabilities. The findings show that mongrel models combining optimization ways with machine literacy ameliorate vaticination delicacy, supporting better decision-making in husbandry. The exploration emphasizes the need for advanced analytics to attack food security and sustainability challenges.

This study [7] applied colorful machine literacy models, including Generalized Neural Network(GRNN), Support Vector Retrogression(SVR), Random Forest(RF), grade Boosting Machine(GBM), and ARIMA, to prognosticate the diurnal noncommercial price of brinjal in 17 requests across Odisha, India. Using data from 1st January 2015 to 31st May 2021, collected from AGMARKNET, the GRNN model outperformed the other models in terms of delicacy. The exploration highlights the potential of advanced neural networks in perfecting agrarian price soothsaying, which can help stakeholders make informed request decisions. This paper [8] explores price soothsaying for essential crops — Tomato, Onion, and Potato(TOP) — in major Indian requests by integrating both price data and exogenous variables like rainfall conditions(rush and temperature). The study compares deep literacy models with traditional styles like ARIMAX and MLR, as well as machine literacy algorithms similar as ANN, SVR, RFR, and XGBoost. Using data from AGMARKNET and rainfall data from NASA POWER, the exploration finds that including rainfall variables improves vaticination delicacy. The study suggests unborn exploration could explore the influence of fresh exogenous factors like news data and social media trends on price soothsaying. This study [9] developed a Crop Price vaticination System using Decision Tree Regression and Random Forest Regression, assaying literal crop price data sourced from data.gov.in. By incorporating fresh input features similar as meteorological parameters and socioprofitable pointers, the models achieved an overall delicacy of 95%, with a peak performance of 97.25% for certain months. The system aims to prop agrarian decision- making by furnishing dependable crop price vaticinations. The authors suggest that unborn work should concentrate on enhancing model robustness, incorporating real- time data, and addressing indigenous variations to further ameliorate vaticination delicacy. This paper [10] introduces the Interaction Regression Model for prognosticating crop yields, particularly fastening on sludge and soybean in three Midwest U.S. states Illinois, Indiana, and Iowa. The model integrates optimization, machine literacy, and agronomic perceptivity, achieving a relative root mean square error of 8% or lower, outperforming several state of the art machine learning algorithms. It identifies crucial terrain- operation relations that affect crop yields, offering both prophetic delicacy and resolvable perceptivity. By anatomizing yield benefactions from rainfall, soil, and operation relations, the model provides agriculturists with precious tools to optimize crop yields grounded on specific environmental conditions. This paper [11] presents a deep literacy frame combining Convolutional Neural Networks(CNNs) and intermittent Neural Networks(RNNs) to prognosticate sludge and soybean yields across 13 countries in the U.S. Corn Belt. Using environmental and operation data from 1980 to 2018, the CNN- RNN model achieved a root- mean- square- error(RMSE) of 9% and 8% of the average yields, outper- forming styles like Random Forest(RF), Deep Completely Connected Neural Networks(DFNN), and LASSO. The model effectively captures time dependences in environmental factors and generalizes prognostications across untested environments without significant delicacy loss. It also reveals how rainfall, soil, and operation practices explain variations in crop yields, offering implicit for broader operation in crop yield studies. The deep neural network (DNN) model used in this [12] study predicts sludge yield using a dataset from the 2018 Syngenta Crop Challenge, taking into account genotype, environmental factors, and their relations. The model outperformed former ways including Lariat, shallow neural networks(SNN), and retrogression trees(RT) and attained a root- mean- forecourt- error(RMSE) of 12% using projected rainfall data. Feature selection decreased the complexity of the input space without appreciably compromising accuracy. The study showed that weather and soil conditions, among other environmental fac- tors, had a bigger influence on yield estimates than genotype, underscoring the significance of environmental data in agricul- tural forecasting. In this [13] study, a hybrid model combining LSTM-RNN (Long Short-Term Memory - Recurrent Neural Network) and Temporal Convolutional Network (TCN) is proposed to predict future crop yields. The model processes historical crop yield data and greenhouse environmental pa- rameters (e.g., CO concentration, temperature, humidity) to capture complex temporal dependencies. By integrating the temporal pattern recognition capabilities of LSTM-RNN and TCN, the approach achieves superior accuracy compared to traditional machine learning and deep learning models. The experimental results demonstrate the hybrid model's effec- tiveness, achieving the lowest mean RMSE across various datasets for greenhouse crop yield prediction.The study of this paper [14] employed a combination of Random Forest (RF), XGBoost, CNN, and a CNN-LSTM-Attention model for crop yield prediction, focusing on data from the critical months of July and August. After performing Exploratory Data Analysis and refining feature selection through corre- lation analysis and Variance Inflation Factor (VIF), RF and XGBoost were used to handle non-linear relationships. A CNN model was utilized for extracting spatial and temporal features, while the CNN-LSTM-Attention model captured deeper tem- poral dependencies, highlighting key features through attention mechanisms. Model performance was evaluated on data from 2014-2019, with validation on 2020 data, using metrics like R^2, RMSE, and MAPE, confirming robust and accurate yield forecasts.This paper [15] explores the use of Random Forest (RF) and Temporal Convolutional Networks (TCN) for crop yield prediction based on satellite data. RF, implemented as a baseline, utilized 500 decision trees to enhance prediction accuracy, leveraging ensemble learning through random fea-

ture sampling at each split. In contrast, TCN was designed to manage sequential data, employing dilated causal convolutions to capture temporal dependencies. Both models were trained separately for each crop type, using a two-year training set and one-year testing set. Cross-validation over multiple runs highlighted the models' effectiveness in predicting crop yields at various stages of the growing season.This paper [16] presents a comprehensive methodology for crop price prediction using several machine learning models, including Linear Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM). The process begins with data collection from sources like Kaggle, followed by preprocessing steps to handle missing values and prepare the dataset. The models are then trained using the train test split method, with their performance evaluated through metrics such as accuracy, precision, recall, and F1-score. The research highlights the system's effectiveness in predicting crop price movements and concludes by discussing future directions for enhancing the models' accuracy and practical application in agricultural pricing.This paper [17] describes a crop price prediction website that employs Decision Tree Regression and Random Forest Regression models to provide accurate forecasts. The Decision Tree Regression method partitions the dataset into leaf nodes based on binary decisions, calculating the average crop price within each node to uncover pricing trends. In contrast, Random Forest Regression enhances prediction accuracy through ensemble learning, constructing multiple decision trees from random data subsets and averaging their outputs. This combined approach effectively addresses complex, non-linear relationships within the data, utilizing historical rainfall and wholesale price data to offer farmers valuable insights for crop selection and financial planning over the next year. The models are trained and evaluated using a 70/30 dataset split, ensuring robust performance and the capability for timely updates as new data becomes available.The [18] proposed methodology for predicting daily agricultural market prices in India integrates a 1-Dimensional Convolutional Neural Network (1D CNN) and a Graph Neural Network (GNN) to enhance prediction accuracy. Utilizing data from the Directorate of Marketing Inspection, the model focuses on daily price and arrival information for crops such as tomatoes and potatoes. The 1D CNN effectively captures temporal changes in weather features, producing compact embeddings that are further processed through fully connected layers for refinement. Simultaneously, the GNN constructs a graph representation where each vertex signifies a mandi, with edges reflecting geographic proximity within a 200 km radius. This structure facilitates tailored predictions based on crop-specific data availability and needs, addressing challenges related to data sparsity and geographical relationships. Overall, this innovative approach significantly improves the accuracy of agricultural market price predictions.

III. Proposed Methodology

This section outlines our structured approach for predicting future crop yield and market price using advanced time series forecasting techniques. We implement four complementary models: ARIMA, RNN, LSTM, and hybrid RNN-LSTM mod- els to ensure robust predictions under varying agricultural and market conditions.

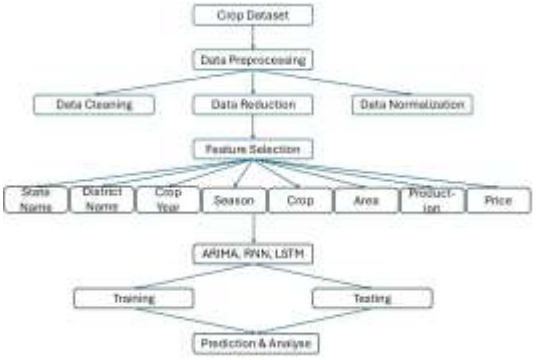

Fig. 2. Overview of the methodology for forecasting agricultural yields and prices.

A. Data Collection and Preprocessing

To train and validate these models, we gathered comprehensive historical datasets covering multiple agricultural seasons and market fluctuations between 2010 and 2024. The datasets include both environmental and economic factors critical for yield and price forecasting.

1) **Crop Yield Data:** We collected district-wise and state-wise historical crop yield data from government agricultural portals and open datasets such as:
   - Ministry of Agriculture databases
   - FAO agricultural statistics
   - Kaggle agricultural yield datasets The dataset includes:
   - Crop type (e.g., wheat, rice, maize)
   - Yield per hectare
   - Cultivation area
   - Season (Kharif, Rabi)
2) **Market Price Data:** For market price prediction, we obtained:
   - Daily and monthly mandi (market) prices from AGMARKNET (Agricultural Marketing Information Network)
   - Minimum Support Price (MSP) records
   - Reports of the state agricultural marketing board
3) **Environmental Data:** To enrich the model with environmental characteristics, we collected:
   - Daily average temperature
   - Rainfall levels
   - Soil type and soil moisture index from Indian Meteorological Department (IMD) and remote sensing data APIs

B. Feature Engineering

1) **Agronomic Feature Construction:** We engineered key agronomic features from the raw agricultural data to capture crop growth patterns and environmental influences:
   - **Growing Degree Days (GDD):** Computed cumulative heat units for each crop season using:

$$GDD = \sum \frac{T_{max} + T_{min}}{2} - T_{base}$$

   where T_{base} is the base temperature threshold for each crop.
   - **Rainfall Index:** Calculated standardized precipitation index (SPI) to quantify rainfall anomalies.
   - **Soil Moisture Index:** Derived from remote sensing soil moisture data for each district and season.
   - **Seasonal Averages:** Computed seasonal mean temperature, cumulative rainfall, and average humidity.
2) **Economic Feature Engineering:** We created additional economic features to enhance market price prediction:
   - **Price Volatility Index:** Measured historical standard deviation of mandi prices for each crop and market.
   - **Lagged Price Features:** Generated lag features (1-month, 3-month) for previous mandi prices.
   - **Minimum Support Price Gap:** Computed the difference between mandi price and MSP to capture price policy effects.
3) **Categorical Encoding:** Categorical variables such as soil type, region, and crop type were encoded using:
   - One-Hot Encoding for soil type and region.
   - Target Encoding for crop type based on average yield.
4) **Dimensionality Reduction:** To mitigate multicollinearity and reduce noise:
   - Applied Principal Component Analysis (PCA) on climate variables (temperature, rainfall, humidity) to extract principal climate patterns explaining 90% variance.
   - The transformed features were calculated as:

$$Z = X_{std} W$$

   where X_{std} is the standardized input matrix and W is the matrix of eigenvectors of the covariance matrix of environmental variables.

C. Crop Yield and Market Price Prediction using ARIMA and RNN

1) **ARIMA Model for Time Series Forecasting:** We implemented an AutoRegressive Integrated Moving Average (ARIMA) model as a baseline statistical approach for univariate time series forecasting of both crop yield and market prices.
   The ARIMA(p, d, q) model is defined as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

   where:
   - y_t is the value at time t
   - p is the order of autoregression
   - d is the degree of differencing
   - q is the order of moving average
   - ϕ_i and θ_j are model coefficients
   - ϵ_t is white noise at time t

   The parameters (p, d, q) were selected by minimizing the Akaike Information Criterion (AIC) and using autocorrelation (ACF) and partial autocorrelation (PACF) plots.
2) **Recurrent Neural Network (RNN) Model Architec- ture:** We also implemented a vanilla Recurrent Neural Network (RNN) to capture sequential dependencies in historical data for price and yield prediction.
   The RNN hidden state update is governed by:

$$h_t = \sigma(W_{xh} x_t + W_{hh} h_{t-1} + b_h) \; y_t = W_{hy} h_t + b_y$$

   where:
   - x_t is the input feature vector at time t
   - h_t is the hidden state at time t
   - y_t is the predicted output at time t
   - W_{xh}, W_{hh}, W_{hy} are learnable weight matrices
   - b_h, b_y are biases
   - σ is the activation function (tanh)
3) **Model Configuration:** The RNN model was configured as follows:
   - Total Two RNN layer first with 80 and second with 60 hidden units
   - Dropout layer with rate = 0.2
   - Dense output layer for yield or price prediction
   - Mean Squared Error (MSE) loss function
   - Adam optimizer with learning rate 0.001
4) **Training Process:**
   - The ARIMA model was trained separately for each crop and region using past 10 years of data
   - The RNN model was trained using time window sequences of 12 time steps (months) for market price prediction and seasonal steps for yield prediction
   - Both models were validated using an 80-20 train-test split and evaluated using RMSE and MAE

D. Crop Yield and Market Price Prediction using LSTM

1) **Long Short-Term Memory (LSTM) Model Architecture:** We implemented a Long Short-Term Memory (LSTM) neural network to model temporal dependencies in historical crop yield and market price data. The LSTM architecture is designed to capture both short-term fluctuations and long-term trends in agricultural time series data. The core LSTM cell is governed by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t * \tanh(C_t)$$

where:

- x_t is the input feature vector at time t (includes rainfall, temperature, soil moisture, lagged price, etc.)
- h_t is the hidden state output at time t
- C_t is the cell state at time t
- f_t, i_t, o_t are the forget, input, and output gates
- W_f, W_i, W_C, W_o are the learnable weight matrices
- b_f, b_i, b_C, b_o are the learnable biases
- σ is the sigmoid activation function

2) **Model Configuration:** The LSTM network consisted of:
   - Two stacked LSTM layers with 80 and 60 hidden units respectively
   - A dropout layer (rate = 0.2) to prevent overfitting
   - A fully connected dense layer for final output prediction (yield or price)
   - Mean Squared Error (MSE) as the loss function
   - Adam optimizer with learning rate 0.001
3) **Input-Output Mapping:** The model was trained to predict:
   - **Crop Yield:** Predicted as continuous output in tons/hectare based on previous season data and environmental factors
   - **Market Price:** Predicted as continuous price per quintal using lagged price, MSP, weather, and demand factors
4) **Training Process:** The LSTM was trained on sliding time windows of 12 months for market price, and seasonal windows for crop yield. Early stopping was applied to avoid overfitting, with validation loss monitored over 20 epochs.

E. Evaluation Framework

We evaluated the performance of the implemented models (ARIMA, RNN, and LSTM) for both crop yield and market price prediction using multiple statistical metrics to ensure robust assessment across different crops and regions:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors between predicted and actual values without considering direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Mean Squared Error (MSE):** Quantifies the average squared difference between predicted and actual values, penalizing larger errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

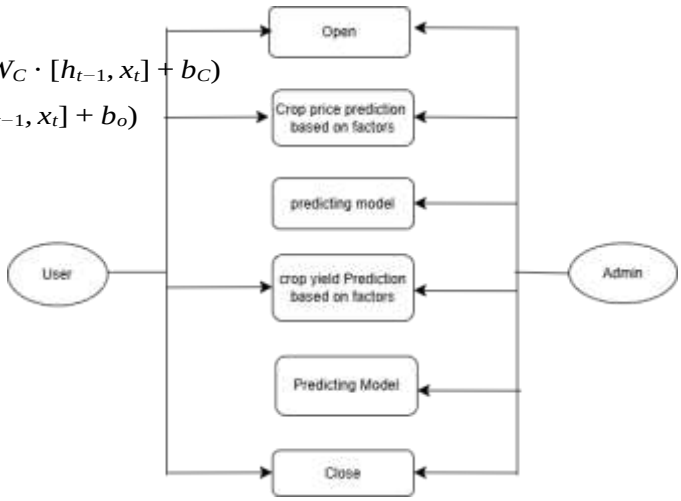

Fig. 3. Use Case Diagram: Interaction between Developer and User for crop price and yield prediction

- **Root Mean Squared Error (RMSE):** Represents the square root of MSE, providing an interpretable measure of prediction error in the original units.

$$RMSE = \sqrt{MSE}$$

- **Mean Absolute Percentage Error (MAPE):** Evaluates prediction accuracy as a percentage, useful for comparing errors across different crops and regions.

$$MAPE = \frac{100\%}{n}\sum_{i=1} \frac{y_i - \hat{y}_i}{y}$$

- **Coefficient of Determination (R^2 Score):** Indicates the proportion of variance in the dependent variable explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

All evaluation metrics were computed for both training and test sets to assess model performance and generalization ability. Cross-validation was performed using k-fold (k=5) to validate consistency across different data splits. Separate evaluations were conducted for each crop type and market region to capture domain-specific prediction quality.

F. Model Performance Visualization

To visually assess the performance of our models, we plotted the actual versus predicted values for both crop yield and market price predictions across selected crops and regions.

- **Crop Yield Prediction Results:** The comparison be- tween actual and predicted crop yields for wheat in Maharashtra demonstrates the effectiveness of the LSTM model over the 10-year test period.
  Fig. 4 shows the comparison between actual yields (green dashed line with circle markers) and predicted yields (red solid line with X markers) for wheat across 10 agricultural years (2015–2024). The model successfully

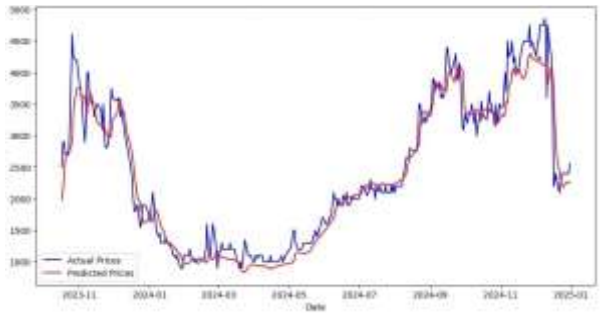

Fig. 4. crop Yield Prediction in Maharashtra (2015–2024) – Actual vs Predicted

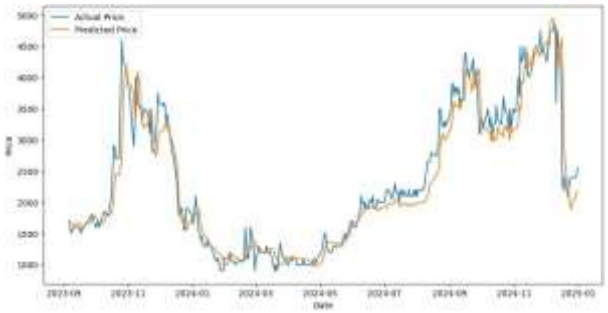

Fig. 5. crop Market Price Prediction (2019–2024) – Actual vs Predicted

captures the overall yield trend, including the drop in 2018 caused by lower rainfall and the yield recovery in 2019.

- **Crop Price Prediction Results:** The RNN model's prediction of average monthly mandi prices for rice in Uttar Pradesh closely follows the actual price trend over the 60-month test period.

  Fig. 5 illustrates the comparison between actual prices (blue solid line) and predicted prices (orange dashed line) for rice market prices between 2019 and 2024. The model demonstrates accurate tracking of price fluctuations, including the price spike during the 2020 pandemic and subsequent stabilization.

These visualizations confirm the ability of the models to generalize between different crops and regions, capturing key patterns and anomalies in both yield and price trends.

NO	Model	Accuracy
1	Arima	78.60
2	RNN	94.50
3	LSTM	96.43

Fig. 6. Result Table

IV. SYSTEM ARCHITECTURE

A. Data Acquisition Layer

This layer manages the collection of agricultural, environmental, and market data from multiple sources:

- **Crop Yield Data Source:** Interfaces with official agricultural databases such as the Ministry of Agriculture, FAO datasets, and Kaggle repositories to collect historical district-wise and state-wise yield data.
- **Market Price Data Source:** Connects to AGMARKNET (Agricultural Marketing Information Network) and state agricultural marketing board reports to retrieve daily and monthly mandi prices, as well as Minimum Support Price (MSP) records.
- **Environmental Data Source:** Accesses weather and soil data from the Indian Meteorological Department (IMD) and remote sensing APIs to gather daily average temper- ature, rainfall, soil moisture, and soil type information.

B. Data Processing Unit

This unit implements the data cleaning and preparation pipeline described in the methodology:

- **Missing Value Handler:** Applies forward-fill for short-term gaps and median imputation for sparse agricultural and environmental records.
- **Outlier Detector:** Utilizes Z-score method with a $\pm 3\sigma$ threshold to detect and handle anomalous values in yield and price data.
- **Normalizer:** Performs min-max scaling for quantitative variables like rainfall, temperature, and price; applies z-score normalization for financial and market features.
- **Temporal Aligner:** Aligns datasets from different sources to a common agricultural calendar, ensuring consistency across crop seasons.

C. Feature Engineering Module

This module extracts predictive features from raw and processed data:

- **Environmental Feature Generator:** Computes derived environmental indicators such as cumulative rainfall, growing degree days, and drought indices.
- **Market Feature Generator:** Calculates moving aver- ages, price volatility measures, and seasonal price indices.
- **Statistical Feature Extractor:** Generates lag features and autocorrelation measures to capture time series dependencies.
- **PCA Transformer:** Applies Principal Component Analy- sis to reduce dimensionality of multivariate environmental and market features while preserving 85% variance.
- **Feature Fusion:** Combines environmental, market, and statistical features into a unified feature set for model input.

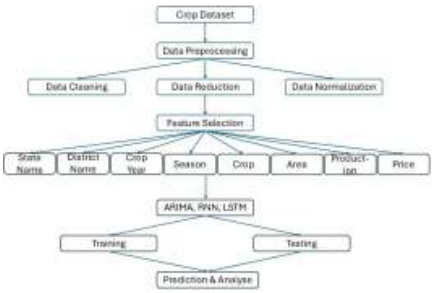

Fig. 7. System Architecture

D. Model Training & Optimization

This component implements the multi-model approach described in the methodology:

- **ARIMA Model Trainer:** Fits AutoRegressive Integrated Moving Average models for univariate yield and price forecasting using historical trends.
- **RNN Model Trainer:** Builds Recurrent Neural Network architecture to learn sequential dependencies in multivariate data for price prediction.
- **LSTM Model Trainer:** Trains Long Short-Term Memory networks for crop yield prediction, integrating environmental and agricultural features.
- **Hyperparameter Optimizer:** Performs grid search and Bayesian optimization for tuning model parameters across ARIMA, RNN, and LSTM models.

E. Evaluation & Monitoring

This component ensures model quality and ongoing monitoring:

- **Model Evaluator:** Computes comprehensive evaluation metrics (MAE, MSE, RMSE, R^2, MAPE) on test data.
- **Performance Visualization Dashboard:** Plots actual vs predicted yields and prices for interpretability.
- **Baseline Comparator:** Benchmarks neural models against traditional models such as linear regression and ARIMA.
- **Model Drift Monitor:** Tracks prediction error over time to detect degradation in performance as new data becomes available.

F. Implementation Technologies

The system is implemented using the following technologies:

- **Core Stack:** Python with Pandas, NumPy, and Scikit-learn for data handling and preprocessing.
- **Deep Learning:** TensorFlow 2.8 and Keras for building and training neural networks (RNN, LSTM).
- **Statistical Modeling:** StatsModels for ARIMA and other classical time series models.
- **Visualization:** Matplotlib and Seaborn for creating data and performance plots.
- **API Development:** Flask or FastAPI for exposing model predictions as web services.

V. FUTURE SCOPE

A. Advanced Model Architectures

Exploring state-of-the-art deep learning approaches for agricultural forecasting.

B. Multi-Region and Multi-Crop Scaling

Expanding the system's applicability across diverse agricultural contexts:

- **Multi-Crop Integration:** Enabling simultaneous prediction across diverse crop varieties (e.g., cereals, pulses, oilseeds, horticultural crops).
- **Cross-Border Data Integration:** Incorporating international agricultural trade data to model price dependencies across neighboring countries.

C. Real-Time Adaptive Learning

Transitioning from static models to continuously evolving predictive systems:

- **Online Learning Frameworks:** Implementing incremen- tal learning pipelines to update models dynamically with incoming mandi prices and seasonal yield data.
- **Meta-Learning Strategies:** Developing models capable of quickly adapting to unseen regions or crops with minimal training data.

These future directions represent significant opportunities to enhance agricultural decision support systems, contributing to sustainable farming practices, market stability, and food security at both regional and national scales.

VI. CONCLUSION

This paper presents a comprehensive predictive system for agricultural forecasting that integrates Long Short-Term Memory (LSTM) networks for crop yield estimation and crop price prediction, alongside traditional time series models such as ARIMA and Recurrent Neural Networks (RNN). The system's modular architecture—spanning data acquisition, preprocess- ing, feature engineering, model training, evaluation, and visualization—demonstrates a scalable and adaptable approach to agricultural data analytics. By incorporating both climatic and economic variables, the proposed framework provides actionable insights to support farmers, policymakers, and supply chain stakeholders. While acknowledging the challenges posed by data sparsity, regional variability, and market volatility, this work represents a significant step toward data-driven agricultural decision support. Future enhancements—including integration of alternative data sources, advanced deep learning architectures, and real-time adaptive learning—hold promise for further improving prediction accuracy and expanding the system's applicability across diverse crops and geographies.

REFERENCES

[1] Girish G P, "Department of Finance,IBSHyderabad, IFHE University," available: https://www.tandfonline.com/doi/epdf/10.1080/23311932.2022.2085717?needAccess=true.
[2] Alejandro Morales,Francisco J. Villalobos" available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1128388/full.

[3] Kiran Moraye,Aruna Pavate,Suyog Nikam,Smit Thakkar, "Atharva Col- lege of Engineering,Mumbai" available:
[4] P Bhasha,Dr. J Suresh Babu,Muniraju Naidu Vadlamudi,Kochumol Abraham,Sanjaya Kumar Sarangi," available: https://www.publishoa. com/index.php/journal/article/view/908/785.
[5] S. Vinson Joshua,A. Selwin Mich Priyadharson,Raju Kannadasan,Arfat Ahmad Khan,Worawat Lawanont,Faizan Ahmed Khan,Ateeq Ur Rehman and Muhammad Junaid Ali, " available:
[6] Seeboli Ghosh Kundu, Anupam Ghosh, Avisek Kundu and Girish G P, A ML-AI ENABLED ENSEMBLE MODEL FORPREDICTING AGRICULTURAL YIELD, available: https://doi.org/10.1080/23311932. 2022.2085717.
[7] Paul RK, Yeasin M, Kumar P, Kumar P, Balasubramanian M, Roy HS, et al. (2022) Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. PLoS ONE 17(7): e0270553. https://doi.org/10.1371/journal.pone.0270553.
[8] Nayak, G.H.H., Alam, M.W., Singh, K.N. et al. Exogenous variable driven deep learning models for improved price forecasting of TOP crops in India. Sci Rep 14, 17203 (2024). https://doi.org/10.1038/ s41598-024-68040-3
[9] Sumit Pandey, Ajay Tanvar, Jayant Lokhande, Pankaj Thorat, Harshada Jadhav, "Crop Price Prediction System Using ML,"https://ijrpr.com/ uploads/V5ISSUE5/IJRPR27405.pdf.
[10] Ansarifar, J., Wang, L. Archontoulis, S.V. An interaction regression model for crop yield prediction. Sci Rep 11, 17754 (2021). https://doi. org/10.1038/s41598-021-97221-7
[11] Saeed Khaki, Lizhi Wang1 and Sotirios V. Archontoulis " A CNN-RNN Framework for Crop Yield Prediction, "https://arxiv.org/pdf/2211. 00974v2
[12] Saeed Khaki, Lizhi Wang , " Crop Yield Prediction Using Deep Neural Networks ," https://arxiv.org/pdf/1902.02860v3
[13] Liyun Gong, Miao Yu, Shouyong Jiang, Vassilis Cutsuridis andSimon Pearson, "Deep Learning Based Prediction on Greenhouse Crop Yield Combined TCN and RNN available:https://www.mdpi.com/1424-8220/ 21/13/4537
[14] Jian Lu, Jian Li, Hongkun Fu, Xuhui Tang, Zhao Liu ORCID,Hui Chen, Yue Sun andXiangyu Ning "Deep Learning for Multi-Source Data-Driven Crop Yield Prediction in Northeast China available:https: //www.mdpi.com/2077-0472/14/6/794
[15] Maria Yli-Heikkila¨, ORCID, Samantha Wittke, ORCID, Markku Lu- otamo ORCID, Eetu Puttonen ORCID, Mika Sulkava, Petri Pellikka ORCID, Janne Heiskanen ORCID and Arto Klami " Scalable Crop Yield Prediction with Sentinel-2 Time Series and Temporal Convolutional Network available:https://www.mdpi.com/2072-4292/14/17/4193
[16] R. Beulah, "A survey on different data mining techniques for crop yield prediction," available:https://doi.org/10.26438/ijcse/v7i1.738744
[17] I. Ahmad, U. Saeed, M. Fahad, A. Ullah, M. Habib-ur-Rahman, A. Ahmad, J. Judge, Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan, J. Indian Soc. Remote Sens., 46 (10) (2018), pp. 1701-1711, 10.1007/s12524-018-0825-8.
[18] A.T.M.S. Ahamed, N.T. Mahmood, N. Hossain, M.T. Kabir, K. Das, F. Rahman, R.M. Rahman, "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh," 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2015 - Proceedings (2015), 10.1109/SNPD.2015.7176185
[19] I. Ali, F. Cawkwell, E. Dwyer, S. Green"Modeling managed grassland biomass estimation by using multitemporal remote sensing data—a ma- chine learning approach," available:https://ieeexplore.ieee.org/document/ 7482764/
[20] https://www.youtube.com/watch?v=7uwa9aPbBRU&list= PLTDARY42LDV7WGmlzZtY-w9pemyPrKNUZ